QuakeSim: Enabling Model Interactions in Solid Earth Science Sensor Webs

PI: Andrea Donnellan

Quarterly Progress Report

December 21, 2006

1. Objective

The objective for QuakeSim under this project (AIST-05) is to integrate both real-time and archival sensor data with high-performance computing applications for data mining and assimilation to improve earthquake forecasts. Ultimately this will lead to mitigation of damage from this natural hazard. We will Develope extensive Geographical Information System-based "Data Grid" services to support ontologically described geographic data archives as well as middleware to support real-time streaming GPS data. This project extends the AIST-02 approach of integrating "Execution Grid" services that are suitable for interacting with high-end computing resources at ARC and JPL.

2. Summary of Accomplishments this Quarter, including the achievement of any milestones

There are no specific milestones for this quarter. The major effort has been administrative with an objective to establish all of the subcontracts with the co-investigator institutions. The PI signed the last of the subcontracts on December 18, 2006. Our plan called for them to be finished by January 2007.

3. Current Progress Description

Web Services. We have two wikis: http://www.crisisgrid.org/cgwiki/ and

www.quakesim.org/wiki. We have been significantly revising the QuakeSim portal. See http://gf2.ucs.indiana.edu:5050/gridsphere/gridsphere. The login is mpierce/mpierce. We have been getting all of our codes (mostly developed under previous AIST projects) under SourceForge Subversion (SVN) version control:

http://crisisgrid.svn.sourceforge.net/viewvc/crisisgrid/.

Our SF project page is http://sourceforge.net/projects/crisisgrid/, but this is not really used for much now except SVN. We revised earlier Perl-based Google/OGC Web Map Server to use Java. This was because we switched developers and is an offshoot of the caching/tiling service (below). This in SVN, above. We deployed and significantly tested our SensorGrid architecture for real time GPS from Scripps. See

http://gf8.ucs.indiana.edu:7474/mythesis/galip-thesis-ppt.ppt.

This was an internal lab presentation by Galip. We have given versions of these to the Scripps GPS group and at Supercomputing 2006. We have been developing the caching/tiling Map Server (which you have seen) and have experimented with making it collaborative.

Data Management. One of the key problems in the effective utilization of available data in earthquake forecasting and related research involves providing an effective interface to InSAR, GPS, and other (high-speed) streamed data. These data must be ingested and stored in a manner effective for selective sharing. In addition, these data must be inter-related with paleoseismic data, fault databases, etc. What is needed is a semantic (ontology-based) federation of these data. The functionality required is for a scientific user to be able to locate specific information they need, for example to use with their modeling or simulation systems. This interface will be an extension of our current Quakesim portal facilities. Once a scientist has found data of interest, and we employ both pull (query) and push (notify the user of things relevant to them), they may wish to visualize it at various levels of abstraction, and in many cases use high performance computing to run their programs on the selected data. This of course involves a program/data interface problem that is to be a part of our web-service based approach to interoperation.

At present, we have a Fault Database System that we developed in the prior funding of this work. Working with the Indiana University (IU) sub-team, we have made an initial assessment of what needs to be done to extend the functionality of this database system to include InSAR and GPS data. A further extension to include more fault data (for all of California) is being planned, involving the University of California, Irvine subteam. We have also initiated the transfer of the management of the database system code from USC graduated Ph.D. students Anne Chen and Sangsoo Sung, to two new Ph.D. students, Rami Elghanmi and Jinwoo Kim. The current system physically resides at the IU and is managed by both USC and IU researchers. We have made progress in extending the search capabilities of this Fault Database System, but have only taken the first step in enriching search and portal request support.

We have also initiated the design of the overall framework for database sharing and interoperation we plan to employ, based upon a federation of ontologies (semantic descriptions of data, connected to web services for accessing those data, Figure 1).

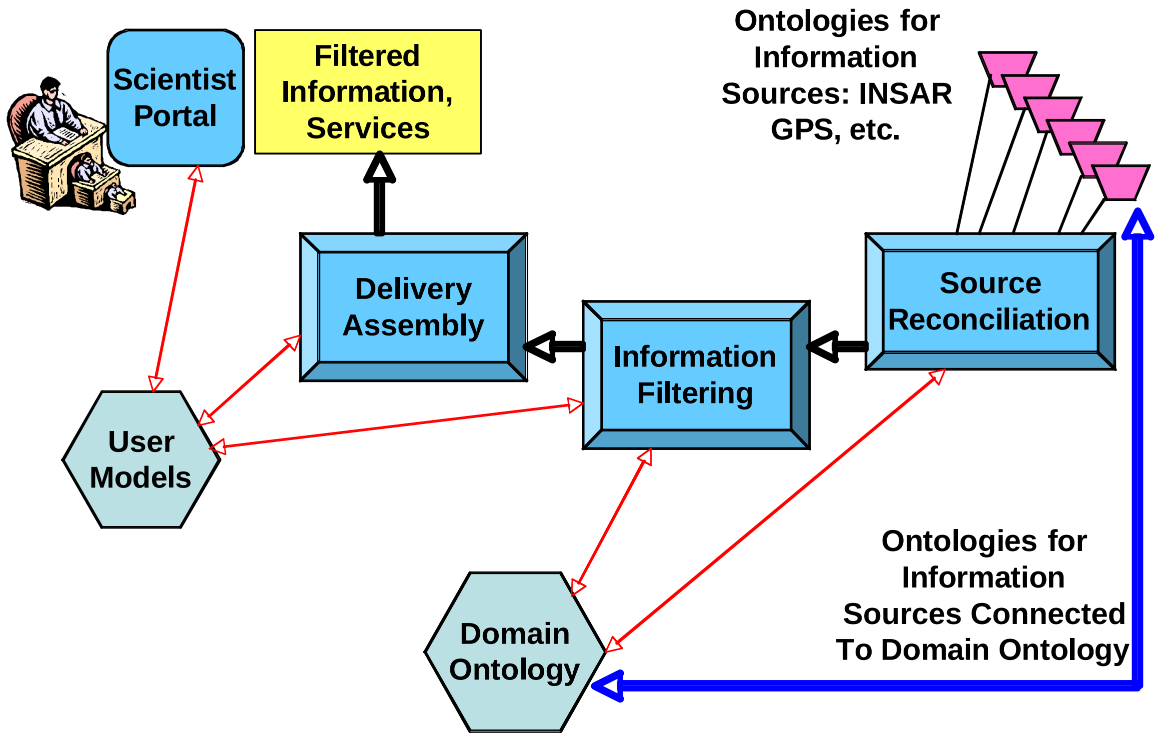

Figure 1. Architecture for information management for QuakeSim

Virtual California. We have used the Columbia computer to develop and optimize the performance of new versions of the Virtual California fault simulation code. Among other improvements, the code is now written in full, C++, object-oriented version with an efficient, fully parallel implementation (in place of the mixed parallel-sequential implementation used heretofore). Our goal with Project Columbia was to use this version of the code, and to optimize it using a number of improvements. We note that the computations carried out during this period utilized a relatively small model having 768 Boundary Elements, which has the effect of producing dominance of the computation by latency issues in the processor-interconnect. With the larger models we now have, such as the more recent 12,288 model ($3*2^{12}$ boundary elements), more computation is done on each processor, so latency issues will be less and less of a problem. Nevertheless, it is important to optimize the performance of the code, so that large models can be computed.

In fact, one of the goals of the current work was to test a new implementation of the failure algorithm, in which load balancing issues are addressed. In the previous version, failure of the elements as the stress threshold is reached typically involved one or two processors doing most of the work, with the other processors standing idle. This newest version of the code is written to redistribute the work in a random shuffling algorithm, so that all processors are used to adjust slip at time of failure during an earthquake.

The latest Virtual California code was ported and tested on the Columbia supercomputer. This code includes the "shuffled" segment distribution among cpus. Previous versions of VC used sequential distribution, so that, for example, if one had 2 cpus, the first cpu would receive the first half of the boundary elements to compute, while the second cpu receives the rest of the elements. With the "shuffled" distribution, the segment set is divided into even and odd subsets.

The shuffled distribution allows us to obtain optimal load balancing. This will be important for the large VC segment sets. A better scaling was also obtained for the current 768 segment set, as can be seen in the Table and the Figure.

On both table and plot, performance data are not shown for 1 and 2 cpus, because Columbia MPI jobs require use of a minimum of 4 cpus. Hence the normalization was done using 4 cpu timing. From results obtained on our "Tribe" cluster at UC Davis, we know that the performance scales almost ideally for 1, 2, and 4 cpu runs. The two plots shown below correspond to sequential (old) and shuffled (new) segment distributions. We obtained extensive simulation data including 32 millions years of VC evolution. All possible cpus configurations (4, 8, 16, 32, and 64) were used as VC simulation results are independent of the cpu count. These data will be analyzed elsewhere.

Table 1. Performance Data. Column 1, number of cpus used. Column 2, time in seconds for sequential adjustment of slip on boundary elements. Column 3, time in seconds for shuffled adjustment of slip on boundary elements.

CPU Count(s)	Sequential (s)	Shuffled
4	417.5	433.5
8	223	255.2
16	177.6	191.5
32	215.3	170.1
64		238.1

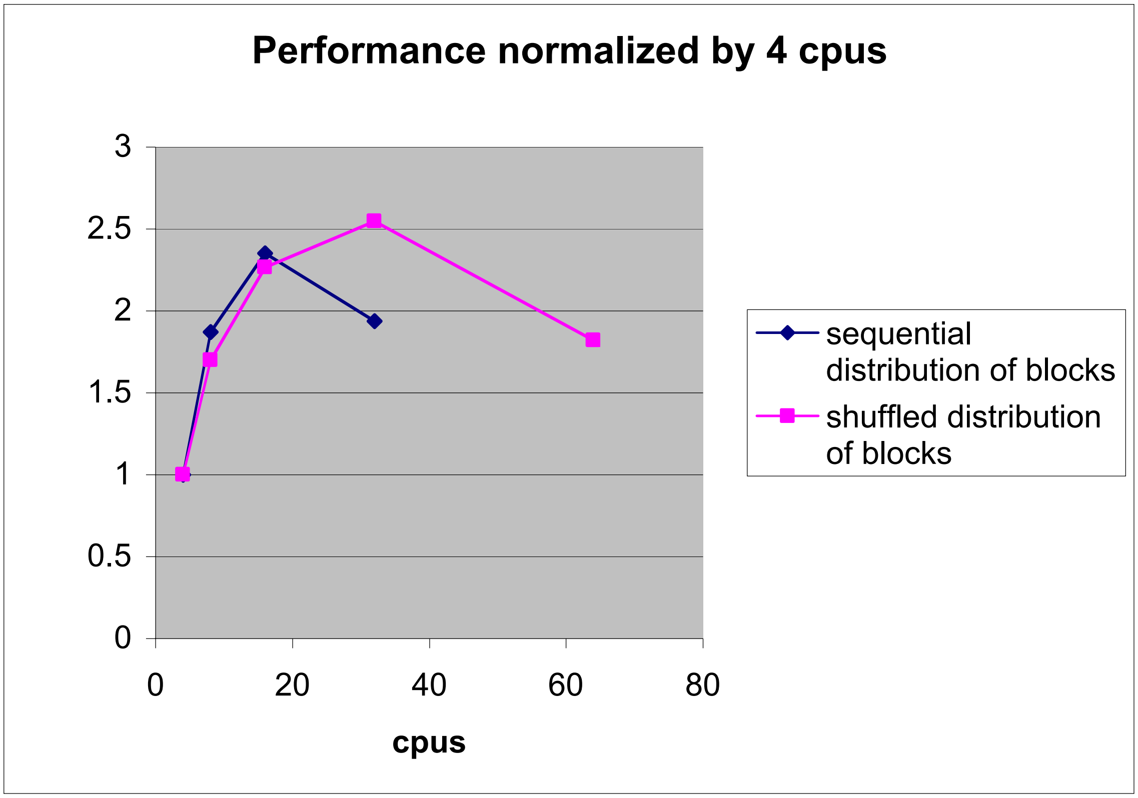

Figure 2. Plot of data in Table 1.

RIPI Method. Our current research progress indicates that the occurrences of large earthquakes in southern California correlate with time intervals where fluctuations in small earthquakes are suppressed relative to the long-term average. We estimate a probability of less than 1% that this coincidence is due to simple random clustering. By studying the systematic fluctuations in smaller sub-regions, using an intervention technique, we attempt to identify spatial locations within the time interval where the correlations are highest. Furthermore, we show that the methods used to obtain these results may be applicable to other parts of the world. In particular, we analyze Japan, Taiwan, and Sumatra and show that the method has potential for forecasting local earthquakes in those regions.

GeoFEST. Extending the QuakeSim environment to high performance computers requires interfacing with a high performance code. In order to test the system, we are porting GeoFEST-4.6 to Project Columbia and are performing testing on various surface deformation problems containing more than 100 million elements. GeoFEST-4.6 has been ported and is running, but we have been experiencing periodic problems with the /nobackup file system in terms of staging and recording simulation test results. The NAS staff has been helpful in addressing this - this is not specific to this task.

Simplex/Inversion. Work has begun on the design of a nonlinear inversion code (an expanded Simplex, perhaps to be called SubSurfer) that includes new deformation model elements. This includes an initial sketch of a main program that appropriately loops over superposed deformation sources, efficiently accounts for parameters and simplex perturbations within each model element, and supports parallel decomposition over the space of observations. Input/output files will be defined as XML items with nested inclusion. Inclusion within the Pyre framework (or re-coding "main" in Python) is being explored at this early stage.

Code has been obtained that will allow better modeling and inversions by supporting deformation within elastic layers with material variation (FORTRAN programs EDGRN/EDCMP, developed at GFZ, Potsdam, Germany; published in Computers & Geosciences v. 29, no. 2, pp. 195-207). I met one of the authors (Francisco Lorenzo) at Fall AGU: he expresses enthusiastic support for our aims.

Internal discussions have begun exploring the use of QuakeTables within modeling codes (including the new inversion code). Initial efforts will design and deploy web services for fault selection and lightweight (perhaps java) tools for assisting users in selecting estimation parameters, modeling constraints, and merging such fault information with earthquake event information.

4. Work Plan for next Reporting Period

We plan to develop initial ontologies to describe the INSAR and GPS data with which we will be working, as we did for the fault data in the current Fault Database System. We will then design database conceptual schemas, utilizing the same platform as the Fault Database System, to handle INSAR and GPS data. We will then construct, perhaps in the next two quarters, a domain ontology for assisting in semantically cross-linking the various databases. Critical to the success of this task is our working with the geoscientists on the project to determine the data they need, how they need to use it. We will begin to assess the cross-correlations and inter-relationships among the fault data, INSAR data, and GPS data – a step towards federating the ontologies and databases for faults, INSAR, and GPS information. We also plan to extend the coverage of the current Fault Database System to all of California over the next two quarters.

Demonstrate capability to run GeoFEST-4.6 on 2048 Lander's simulation problem. Hope to release through OpenChannel a major GeoFEST release next quarter, with buoyancy and improved refinement methods for viscoelastic time steps (currently in debugging mode). We are working to revise the GeoFEST portlets and use Condor-G to submit GeoFEST jobs to TeraGrid machines.

Continue optimization of Virtual California in C++ MPI format. Conduct scaling studies on Columbia computer using new codes, interface with surface deformation codes, develop new stress green's function codes for dipping fault sources, develop next generation models for VC 2006 (topology and friction laws). Generate new types of output with depth dependent slip. Continue work to optimize and refine RIPI models. Develop methods based on intervention to locate epicenters of future earthquakes.

We will begin work on inversion application and it's integration with the web services portions of QuakeSim. We will work with IU and USC to design web service and portlet tools to bring faults and event information into inversions for subsurface processes. We will begin planning changes for QuakeTables interface to enhance utility by modelers.

We plan next quarter to start to integrate the Map Server with maps of California and contour plots of GeoFEST output. See

http://www.crisisgrid.org/cgwiki/images/4/49/IGIC_Oct27_2006_Final.ppt,

which we presented to Indiana Geographic Information Council and several other groups around the state (Department of Environmental Management, State Department of Homeland Security). Schedule Status (include any slippages or accelerations in schedule and note the cause).

5. Delays/Problems Experienced

There are no delays or problems directly related to this project.

Minor delays have occurred in terms of resolving file system issues with /nobackup system on Columbia and access to 2048 system.

6. Corrective Actions/Recovery Plan

Working with NAS staff to address system issues.

7. Technology Readiness Level Assessment

The current TRL of the project is 3. The concept of developing Web Services for high-performance computers has been formulated but not demonstrated for this project.

8. Publications and Presentations

Grant, L., Donnellan, A., McLeod, D., Pierce, M., Fox, G., Chen, Y., Gould, M., Sung, S., and Rundle, P., "A Web-Service Based Universal Approach to Heterogeneous Fault Databases", *Computing in Science and Engineering Special Issue on Multi-Physics Modeling*, 2006.

Chung, S., Jun, J, and McLeod, D., "A Web-Based Novel Term Similarity Framework for Ontology Learning", *Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2006.

Sung, S. and McLeod, D., "Ontology-Driven Semantic Matches between Database Schemas", *Proceedings of International Workshop on Database Interoperability*, Atlanta GA, 2006.

Norton, C.D., J.W. Parker, G.A. Lyzenga, M.T. Glasscoe, A. Donnellan, Design and Performance Optimization of GeoFEST for Adaptive Geophysical Modeling on High Performance Computers, AGU Fall Meeting, San Francisco, CA, 2006.

Van Aalsburg, J., L. Grant, J. Holliday, and J.B. Rundle, Assimilating Data Into Virtual California Models by a Method of Data Scoring, presented at the Fall 2006 AGU meeting, San Francisco, CA, 2006.

Rundle, P.B., J.B. Rundle, G. Yakovlev, J. Fernandez, R. Shcherbakov, D.L. Turcotte, A. Donnellan, N. Field, L. Grant, K.F. Tiampo, J. Van Aalsburg, W. Klein and L.H.

Kellogg, Simulating the San Andreas Plate Boundary System: Progress and Prospects, presented at the Fall 2006 AGU meeting, San Francisco, CA, 2006.

Validation of Quasi-Static Methods for Southern California Deformation, Jay Parker, Gregory Lyzenga, Cinzia Zuffada, Charles Norton, Margaret Glasscoe, Andrea Donnellan, AGU Fall Meeting 2006, San Francisco, CA.

Parker, J., G. Lyzenga, C. Norton, C. Zuffada, M. Glasscoe, J. Lou, A. Donnellan, Geophysical Finite Element Simulation Tool (GeoFEST): algorithms and validation for quasistatic regional faulted crust problems, Pure and Applied Geophysics special issue, submitted.

9. Additional Comments or Pertinent Information

None